BioMedGPT: Open Multimodal Generative Pre-trained Transformer for BioMedicine

Abstract

Foundation models (FMs) have exhibited remarkable performance across a wide range of downstream tasks in many domains. Nevertheless, general-purpose FMs often face challenges when confronted with domain-specific problems, due to their limited access to the proprietary training data in a particular domain. In biomedicine, there are various biological modalities, such as molecules, proteins, and cells, which are encoded by the language of life and exhibit significant modality gaps with human natural language. In this paper, we introduce BioMedGPT, an open multimodal generative pre-trained transformer (GPT) for biomedicine, to bridge the gap between the language of life and human natural language. BioMedGPT allows users to easily ``communicate'' with diverse biological modalities through free text, which is the first of its kind. BioMedGPT aligns different biological modalities with natural language via a large generative language model, namely, BioMedGPT-LM. We publish BioMedGPT-10B, which unifies the feature spaces of molecules, proteins, and natural language via encoding and alignment. Through fine-tuning, BioMedGPT-10B outperforms or is on par with human and significantly larger general-purpose foundation models on the biomedical QA task. It also demonstrates promising performance in the molecule QA and protein QA tasks, which could greatly accelerate the discovery of new drugs and therapeutic targets. In addition, BioMedGPT-LM-7B is the first large generative language model based on Llama2 in the biomedical domain, therefore is commercial friendly. Both BioMedGPT-10B and BioMedGPT-LM-7B are open-sourced to the research community. In addition, we publish the datasets that are meticulously curated for the alignment of multi-modalities, i.e., PubChemQA and UniProtQA. All the models, codes, and datasets are available at \url{https://github.com/PharMolix/OpenBioMed}.